Content Based Music Genre Classification using Temporal and Spectral Features

Mayank Bansal (20111032)
Gaurav Tank (20111407)
Harshwardhan Pratap Singh (20111410)
Debanjan Chatterjee (xxxxxx)

• Motivation:

Music is heard by everyone. It propagates feelings from one to another. Today's music corpus is very heterogeneous in nature. Everybody has their own music tests. Mainly because of diversity of composers and musicians. However, many music streaming industries like pandora, spotify, google use different state of the art learning algorithms to find similarities and patterns between tracks; and based on that, tracks are binned into different classes known as 'Genres' of the tracks. Based on the genres one listen to, it recommends related audio tracks and songs.

As music listeners, we are motivated to get insight into some of the existing music classification techniques and also to try out experiments using different machine learning algorithms and see if we could improve some existing techniques.

• Method:

In machine learning, individual examples have different features and based on that, practitioners apply some intuitions and domain knowledge and tries to find pattern in the data. Here, our dataset will consists of labelled audio tracks. where label represents the genre of the songs/tracks. So, we will make use of different supervised learning techniques to find out patterns and classify the songs/tracks into different categories.

• Intended Experiments:

Audio data is represented by a digital signal. The audio data is already arranged in time series fashion. However, we can't apply end to end learning techniques due to its sampling rate. The standard audio signal is encoded with sampling rate around 44100Hz. Suppose we are able to identify genre after listening to 10 seconds, then it will have around (10 * 44100 = 441000) samples and processing such sequence is not feasible.

We can make use of different existing and well-known feature extracting techniques like discrete fourier transform (DFT), zero crossing rate (ZCR), Mel Frequency Cepstral Coefficients (MFCC), Chroma features and many more. In most of the multimedia processing literatures, these methods do exist so, we do hope that these techniques will help us in the feature extraction task. We are planning to use short-time fourier transform with appropriate window-size and hop-length to create different spectrums for different tracks on different timestamps. This will create temporal sequence but with very less timestamps but this will increase the dimensionality of feature for a timestamp. But, this way, we can use parallel processing on individual timestamps.

After extracting the proper features, next step is to learn from these features. Recurrent Neural Networks are most suitable here due to their memorization capabilities. So we will try out different architectures having LSTM and GRU units. Our input unit will take high dimensional data and our last layer will have softmax units because this will be a simple classification task. And we will experiment will number of hidden units and try out different activation functions.

• Dataset:

There are many datasets are available on the internet. We are planning to use some pre-existing dataset from the conferences' websites like ISMIR.

• Future Work:

Songs do not only have auditary data. They have other metadata associated with that like lyrics in the form of text data, artists' cover images in the form of image data. So, we can train new additional models on these forms of data. and then we can add those individual models into our system to build an ensemble learning model. That might reduce the classification/prediction errors from individual models.